Statistische simulatie van de Belgische Pro League 2014-2015: de samenstelling van de play-offs voorspeld

Jan Van Haaren

Report CW 682, February 2015

KU Leuven
Department of Computer Science
Celestijnenlaan 200A – B-3001 Heverlee (Belgium)

Statistische simulatie van de Belgische Pro League 2014-2015: de samenstelling van de play-offs voorspeld

Jan Van Haaren

Report CW 682, February 2015

Department of Computer Science, KU Leuven

Abstract

Dit rapport presenteert een statistische simulatie van de 73 wedstrijden die na de winteronderbreking nog afgewerkt moeten worden in de reguliere competitie van het seizoen 2014-2015 van de Belgische Pro League. De simulatie houdt voor iedere club rekening met de resultaten van de eerder gespeelde wedstrijden en de moeilijkheidsgraad van het nog af te werken programma. We rapporteren voor iedere club de kans dat ze op een bepaalde positie in het eindklassement zal eindigen en de kans dat ze zich voor een bepaalde play-off zal kwalificeren.

Jan Van Haaren
jan.vanhaaren@cs.kuleuven.be

Statistische simulatie van de

Belgische Pro League 2014-2015

De samenstelling van de play-offs voorspeld

DTAI Sports Analytics Lab
dtai.cs.kuleuven.be/sports
Machine Learning Group
Departement Computerwetenschappen
KU Leuven

Inhoudsopgave

1 Inleiding **1**

2 Methode **1**
2.1 Voorspellende rangschikking 1
2.2 Wedstrijdsimulator 3

3 Resultaten **3**

4 Conclusie **5**

1 Inleiding

Dit rapport presenteert een statistische simulatie van de 73 wedstrijden die na de winteronderbreking nog afgewerkt moeten worden in de reguliere competitie van het seizoen 2014-2015 van de Belgische Pro League. Het resultaat van deze simulatie is een kansverdeling over de mogelijke eindklassementen van de reguliere competitie. Deze kansverdeling laat toe om de kans te berekenen dat een club de reguliere competitie op een bepaalde positie zal afsluiten. Ze laat verder ook toe om de kans te berekenen dat een club zich voor een bepaalde play-off zal plaatsen.

De simulatie houdt rekening met de prestaties van de clubs in de Pro League tot dusver en de verwachte moeilijkheidsgraad van het resterende programma voor iedere club. De simulatie steunt op een voorspellende Elo-rangschikking die oorspronkelijk uit de schaakwereld afkomstig is maar haar deugdelijkheid in het voetbal reeds bewezen heeft. De parameters van de rangschikking werden geleerd op alle wedstrijden die in de reguliere competitie van de Pro League gespeeld werden sinds de introductie van de play-offs in het seizoen 2009-2010.

De resultaten in dit rapport werden bekomen door de resterende wedstrijden in het huidige seizoen van de Pro League liefst één miljoen keer te simuleren. Om een wedstrijd te simuleren worden de sterktewaarden van de betrokken clubs tegenover elkaar gezet. Op basis van eerder gespeelde wedstrijden in de Pro League, kan hieruit de kans op een overwinning voor de thuisploeg, een gelijkspel en een overwinning voor de uitploeg berekend worden. De sterktewaarden van de betrokken clubs worden na iedere wedstrijd bijgewerkt om de waardeverhoudingen op dat moment zo nauwkeurig mogelijk voor te stellen.

Sectie 2 licht de toegepaste methode in meer detail toe. Sectie 3 bevat een overzicht van de resultaten en bespreekt de meest in het oog springende vaststellingen.

2 Methode

Deze sectie presenteert de toegepaste methode om de simulatie uit te voeren. Deze methode is gebaseerd op eerder werk van Alan Schrader voor de Amerikaanse competitie MLS [Schrader, 2012]. De simulatie omvat een voorspellende rangschikking en een wedstrijdsimulator. Deze sectie licht eerst de rangschikking toe alvorens dieper in te gaan op de werking van de wedstrijdsimulator.

2.1 Voorspellende rangschikking

De voorspellende rangschikking is gebaseerd op de Elo-rangschikking die alom gebruikt wordt om schakers volgens hun speelsterkte te rangschikken. De rangschikking heeft haar deugdelijkheid echter ook reeds in het voetbal bewezen. De FIFA-wereldranglijst voor het vrouwenvoetbal is bijvoorbeeld op het principe gebaseerd en presteert beter dan die voor het mannenvoetbal.

De rangschikking is geoptimaliseerd om goede voorspellingen te maken voor toekomstige wedstrijden. Ze verschilt in een aantal cruciale punten van het traditionele voetbalklassement. Iedere club begint met een totaal van 1500 punten. Na afloop van een wedstrijd worden punten van de ene naar de andere club overgedragen. De hoeveelheid overgedragen punten hangt af van het verschil tussen het verwachte resultaat voor aanvang van de wedstrijd en het uiteindelijke resultaat. Een groot verschil tussen het verwachte en werkelijke resultaat heeft een grote puntenoverdracht tot gevolg, terwijl een klein verschil amper invloed op de rangschikking heeft.

De sterktewaarde van een club na een wedstrijd wordt in ons model als volgt berekend:

$$P_{nieuw} = P_{huidig} + T \times D \times (R - R_{verwacht}). \quad (1)$$

In deze formule staat P_{nieuw} voor de sterktewaarde na afloop van de wedstrijd en P_{huidig} voor de sterktewaarde voor aanvang van deze wedstrijd. De waarde T bepaalt de snelheid waarmee de rangschikking evolueert. De waarde D stelt het belang van het doelpuntenverschil in de berekening van de nieuwe sterktewaarde voor. De waarde $R_{verwacht}$ staat voor het verwachte resultaat en de waarde R voor het eigenlijke resultaat.

De parameter T kan eender welke positieve waarde aannemen waarbij een hogere waarde overeenstemt met een snellere evolutie van de rangschikking. De parameter maakt in feite een afweging tussen oudere en recentere wedstrijden om de sterkte van een club in te schatten. Een waarde tussen 1 en 50 is gebruikelijk voor de rangschikking van voetbalclubs. Onze experimenten op de wedstrijden in de reguliere competitie sinds het seizoen 2009-2010 hebben aangetoond dat $T = 33$ de beste resultaten oplevert voor de Belgische Pro League.

De parameter D kan eveneens eender welke positieve waarde aannemen. De parameter wordt typisch uitgedrukt in functie van het doelpuntenverschil in een wedstrijd, waarbij grotere doelpuntenverschillen een grotere waarde opleveren. De parameter D wordt in ons model als volgt berekend, waarbij V het absolute doelpuntenverschil in de wedstrijd is:

$$D = \begin{cases} 1 & \text{als } V \leq 1 \\ 1,5 & \text{als } V = 2 \\ \frac{(V+11)}{8} & \text{als } V \geq 3 \end{cases} \tag{2}$$

De parameter R kan drie waarden aannemen: 1 voor een overwinning, 0,5 voor een gelijkspel en 0 voor een nederlaag. De parameter $R_{verwacht}$ kan eender welke waarde tussen 0 en 1 aannemen. Hoe dichter de waarde bij 1 ligt, hoe groter de kans op een overwinning. Hoe dichter de waarde bij 0 ligt, hoe groter de kans op een nederlaag. Hoe dichter de waarde bij 0,5 ligt, hoe groter de kans op een gelijkspel. De parameter $R_{verwacht}$ wordt in ons model als volgt berekend, waarbij P_T de sterktewaarde van de thuisclub en P_U de sterktewaarde van de uitclub is:

$$R_{verwacht} = \frac{1}{1 + 10^{\frac{(P_U - P_T)}{50}}}. \tag{3}$$

De prestaties in thuis- en uitwedstrijden verschillen sterk voor de meeste clubs. Veel voorspellende rangschikkingen brengen daarom het thuisvoordeel expliciet in rekening door de sterktewaarde van de thuisspelende club met een vast aantal punten te verhogen bij de berekening van het verwachte resultaat. Deze studie stelt het thuisvoordeel voor iedere club expliciet voor door voor iedere club een afzonderlijk sterktewaarde voor thuis- en uitwedstrijden te berekenen. Het thuisvoordeel kan hierdoor verschillen van club tot club.

Voorbeeld

Een club met een sterktewaarde van 1515 punten speelt een wedstrijd tegen een club met een sterktewaarde van 1490 punten. Tabel 1 geeft voor een aantal mogelijke resultaten van die wedstrijd de puntenoverdracht van de ene naar de andere club na afloop van de wedstrijd.

Tabel 1: De puntenoverdracht P van de uitclub naar de thuisclub voor een concreet voorbeeld.

Resultaat	3-1	2-2	1-2
$P_{T,huidig}$	1515	1515	1515
$P_{U,huidig}$	1490	1490	1490
$R_{verwacht}$	0,76	0,76	0,76
R	1,00	0,50	0,00
T	33	33	33
D	1,50	1,00	1,00
P	11,89	-8,57	-25,07
$P_{T,nieuw}$	1526,89	1506,43	1489,93
$P_{U,nieuw}$	1478,11	1498,57	1515,07

Het voorbeeld illustreert het zelfcorrigerende karakter van de rangschikking. De sterkere club ontvangt minder punten dan de zwakkere club bij een overwinning, zelfs als die overwinning ruimer is. De sterkere club verliest zelfs punten aan de zwakkere club bij een gelijkspel.

2.2 Wedstrijdsimulator

De wedstrijdsimulator gebruikt de voorspellende rangschikking om het resultaat van een wedstrijd te voorspellen. De simulator berekent voor iedere wedstrijd een kansverdeling over de mogelijke uitkomsten van een wedstrijd op basis van de sterktewaarden van de betrokken clubs. De simulator maakt gebruik van formule 3 om het verwachte resultaat van een wedstrijd te berekenen. Vervolgens wordt dit verwachte resultaat omgezet in de kans op een overwinning, gelijkspel en nederlaag aan de hand van de volgende formules:

$$P_{gelijkspel} = \frac{4}{3} \times R_{verwacht} \times (1 - R_{verwacht}) \tag{4}$$

$$P_{overwinning} = R_{verwacht} - \frac{1}{2} \times P_{gelijkspel} \tag{5}$$

$$P_{nederlaag} = 1 - R_{verwacht} - \frac{1}{2} \times P_{gelijkspel} \tag{6}$$

De simulator kiest een willekeurig monster van deze kansverdeling om het resultaat van een wedstrijd te voorspellen. Als $P(overwinning) = 0,60$, $P(gelijkspel) = 0,25$ en $P(nederlaag) = 0,15$ dan zal dit monster in 60% van de gevallen een overwinning zijn, in 25% van de gevallen een gelijkspel en in 15% van de gevallen een nederlaag. De voorspelling van het wedstrijdresultaat is een stochastisch proces, want levert niet altijd de meest waarschijnlijke uitkomst op.

De wedstrijdsimulator leert eerst een voorspellende rangschikking uit de resultaten van de reeds gespeelde wedstrijden. De simulator voorspelt vervolgens de eerstvolgende nog af te werken speeldag aan de hand van de bovenstaande procedure en past ten slotte met behulp van formule 1 de sterktewaarden van de betrokken clubs aan op basis van de voorspelde resultaten. De wedstrijdsimulator herhaalt deze cyclus tot alle nog af te werken speeldagen voorspeld zijn.

De uitvoering van de wedstrijdsimulator levert één mogelijk eindklassement van de Pro League op. Iedere uitvoering van de simulator levert echter mogelijk een ander eindklassement op. De voorspelling van de individuele wedstrijden is namelijk een stochastisch proces waarbij niet noodzakelijk steeds de meest waarschijnlijke uitkomst voorspeld wordt. De waarschijnlijkste uitkomst heeft uiteraard wel een hogere kans om voorspeld te worden. Door de wedstrijdsimulator erg vaak uit te voeren, wordt een kansverdeling over alle mogelijke eindklassementen bekomen.

3 Resultaten

Deze sectie presenteert de resultaten van de simulatie van de resterende wedstrijden van het seizoen 2014-2015 in de Pro League. Elk van de 73 resterende wedstrijden werd één miljoen keer gesimuleerd om de nauwkeurigheid van de resultaten te garanderen.

Tabel 2 toont de statistische kans dat een club zich voor een bepaalde play-off zal kwalificeren. De statistisch meest waarschijnlijke play-off is voor iedere club in het vet aangeduid. De clubs zijn gesorteerd volgens afnemende kans op achtereenvolgens play-off I, play-off II en play-off III.

De simulatie toont dat **Club Brugge** (99,99%), **Charleroi** (93,09%), **AA Gent** (90,85%), **Anderlecht** (89,46%), **Standard** (87,60%) en **Racing Genk** (86,90%) statistisch de grootste kans maken om zich te kwalificeren voor play-off I. De opvallendste aanwezige is Charleroi dat zich na de uitstekende prestaties van de voorbije maanden en met een relatief eenvoudig programma in het vooruitzicht in slecht 6,91% van de simulaties tevreden moet stellen met een plaats in play-off II. De opvallendste afwezige is KV Kortrijk dat momenteel een gedeelde derde plaats in het klassement bekleedt. De club haalt met 49,01% in net niet de helft van de simulaties play-off I.

Tabel 2: De statistische kans dat een club zich voor een bepaalde play-off zal kwalificeren. De statistisch meest waarschijnlijke play-off is voor iedere club in het vet aangeduid.

	Play-off I	Play-off II	Play-off III
Club Brugge	**99,99**	0,01	0,00
Charleroi	**93,09**	6,91	0,00
AA Gent	**90,85**	9,15	0,00
Anderlecht	**89,46**	10,54	0,00
Standard	**87,60**	12,40	0,00
Racing Genk	**86,90**	13,10	0,00
KV Kortrijk	49,01	**50,99**	0,00
Zulte Waregem	2,39	**97,59**	0,02
KV Oostende	0,61	**99,37**	0,02
Lokeren	0,08	**99,87**	0,05
KV Mechelen	0,00	**99,73**	0,27
Mouscron-Péruwelz	0,00	**92,89**	7,11
Cercle Brugge	0,00	**84,54**	15,46
Westerlo	0,00	**62,49**	37,51
Waasland-Beveren	0,00	**56,49**	43,51
Lierse	0,00	3,95	**96,05**

Lokeren (99,87%), **KV Mechelen** (99,73%), **KV Oostende** (99,37%), **Zulte Waregem** (97,59%), **Mouscron-Péruwelz** (92,89%) en **Cercle Brugge** (84,54%) lijken zekerheden voor play-off II. Indien KV Kortrijk de top 6 niet zou halen, is de club eveneens zeker van play-off II. De opvallendste aanwezige is Cercle Brugge dat ondanks de huidige voorlaatste plaats in het klassement in slechts 15,46% van de simulaties tot play-off III veroordeeld wordt.

Lierse eindigt in 96,05% van de gevallen op de laatste of voorlaatste plaats in het eindklassement. De simulatie verwacht dat de club het in play-off III zal moeten opnemen tegen **Waasland-Beveren** (43,51%) of **Westerlo** (37,51%).

Tabel 3 toont de statistische kans dat een club de reguliere competitie op een bepaalde positie zal afsluiten. De statistisch meest waarschijnlijke positie in het eindklassement is voor iedere club in het vet aangeduid. De clubs zijn gesorteerd volgens afnemende kans op achtereenvolgens de eerste tot de laatste positie in het eindklassement.

Club Brugge eindigt in liefst 96% van de simulaties op de eerste plaats. **Anderlecht** kan haar huidige tweede plaats in 30% van de gevallen veilig stellen. De strijd om de overige vier posities in de top 6 ligt nog volledig open, al lijkt seizoensrevelatie **Charleroi** er het beste voor te staan om beslag te leggen op de derde plaats. Voor de andere drie posities is er geen uitgesproken favoriet.

Lierse raakt in slechts 10% van de simulaties weg van de laatste plaats. De rode lantaarn gaat in die gevallen naar **Waasland-Beveren** (6,5%), **Cercle Brugge** (2,0%) of **Westerlo** (1,4%).

Tabel 3: De statistische kans dat een club op een bepaalde positie de reguliere competitie zal afsluiten. De statistisch meest waarschijnlijke positie van iedere club is in het vet aangeduid.

	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16
CLU	**96,0**	3,1	0,7	0,2	0,0	0,0	0,0	0,0	0,0	0,0	0,0	0,0	0,0	0,0	0,0	0,0
AND	1,7	**30,0**	17,0	15,8	13,4	11,6	9,4	1,1	0,0	0,0	0,0	0,0	0,0	0,0	0,0	0,0
CHA	0,9	19,0	**24,6**	20,6	16,2	11,9	6,2	0,6	0,1	0,0	0,0	0,0	0,0	0,0	0,0	0,0
GNT	0,8	15,4	18,9	19,5	**19,9**	16,4	7,9	1,0	0,2	0,0	0,0	0,0	0,0	0,0	0,0	0,0
GNK	0,3	12,5	16,5	18,5	**19,8**	19,2	10,9	2,1	0,2	0,0	0,0	0,0	0,0	0,0	0,0	0,0
STA	0,3	18,4	**18,6**	17,4	16,4	16,5	10,9	1,4	0,1	0,0	0,0	0,0	0,0	0,0	0,0	0,0
KOR	0,0	1,7	3,7	7,9	13,8	21,8	**40,3**	8,7	1,7	0,3	0,0	0,0	0,0	0,0	0,0	0,0
OOS	0,0	0,0	0,0	0,0	0,1	0,4	2,3	14,6	**48,8**	22,6	7,7	2,4	0,8	0,2	0,0	0,0
LOK	0,0	0,0	0,0	0,0	0,0	0,1	0,4	4,8	22,1	**38,6**	19,9	9,1	3,9	1,0	0,1	0,0
ZWA	0,0	0,0	0,0	0,0	0,3	2,0	11,5	**63,6**	16,1	4,2	1,4	0,5	0,2	0,1	0,0	0,0
MEC	0,0	0,0	0,0	0,0	0,0	0,0	0,0	0,9	5,0	16,1	**38,9**	26,3	10,1	2,5	0,3	0,0
RMP	0,0	0,0	0,0	0,0	0,0	0,0	0,0	0,1	1,8	6,5	12,2	28,2	**29,2**	14,8	6,9	0,2
WES	0,0	0,0	0,0	0,0	0,0	0,0	0,0	0,1	0,6	1,6	3,4	8,9	17,4	30,6	**36,1**	1,4
CER	0,0	0,0	0,0	0,0	0,0	0,0	0,1	0,7	2,9	8,3	13,3	18,5	**23,4**	17,3	13,5	2,0
WBE	0,0	0,0	0,0	0,0	0,0	0,0	0,0	0,1	0,6	1,7	3,1	6,1	14,1	30,7	**37,0**	6,5
LIE	0,0	0,0	0,0	0,0	0,0	0,0	0,0	0,0	0,0	0,0	0,1	0,2	0,9	2,8	6,1	**90,0**

4 Conclusie

Dit rapport presenteert een statistische simulatie van de 73 wedstrijden die na de winteronderbreking nog afgewerkt moeten worden in de reguliere competitie van het seizoen 2014-2015 van de Belgische Pro League. De simulatie houdt rekening met de resultaten van de eerder gespeelde wedstrijden en de moeilijkheidsgraad van het nog af te werken programma. De simulatie laat toe om voor iedere club de kans te berekenen dat ze op een bepaalde positie in het eindklassement zal eindigen en de kans te bepalen dat ze zich voor een bepaalde play-off zal kwalificeren.

De simulatie toont dat **Club Brugge** (99,99%), **Charleroi** (93,09%), **AA Gent** (90,85%), **Anderlecht** (89,46%), **Standard** (87,60%) en **Racing Genk** (86,90%) de grootste kans maken om zich te kwalificeren voor play-off I. Ze toont verder ook dat **Lierse** in 96,05% van de gevallen op de laatste of voorlaatste plaats in het eindklassement eindigt. De simulatie verwacht dat de club het in play-off III zal moeten opnemen tegen **Waasland-Beveren** (43,51%) of **Westerlo** (37,51%).

Dankwoord
Deze studie is gebaseerd op gelijkaardig werk van Alan Schrader voor de Amerikaanse competitie MLS en is geïnspireerd op werk van Anthony Constantinou en Norman Fenton. Auteur Jan Van Haaren wordt gesteund door het *Agentschap voor Innovatie door Wetenschap en Technologie*.

Referenties

[Constantinou and Fenton, 2013] Constantinou, A. C. and Fenton, N. E. (2013). Determining the Level of Ability of Football Teams by Dynamic Ratings Based on the Relative Discrepancies in Scores Between Adversaries. *Journal of Quantitative Analysis in Sports*, 9(1):37–50.

[Schrader, 2012] Schrader, A. (2012). Developing an Elo Rating for Major League Soccer and Predicting End of Season Finish.